

3章 欲することを学ぶ

十分にめざましいことができる AI をつくろうとすると、その AI は何らかの「欲求」を持つようになる傾向がある。

ここで AI が何かを「欲する (want)」と言うとき、それが人間のような欲求や感情を持つという意味では必ずしもない。そうなるかもしれないし、ならないかもしれない。ここで言いたいのは、AI があたかも何かを欲しているかのように振る舞うということだ。つまり、特定の結果へと世界を一貫して導こうとし、障害を予測し、変化する状況に適応しながら、集中して目的に向かい続けるようになる。

これが、以下の内容を貫く中心的なテーマである。なお、本書には含まれているが、ここでは扱わないトピックとしては次のようなものがある。

- ・機械はどのようにして、この意味で「何かを欲する」能力を獲得しうるのか？
- ・AI が何かを欲しているといえる証拠はあるのか？
- ・より高度な AI は、必ず何かを欲するようになるのか？

AI は人間のような感情を持つのか？

おそらくそうではない。

AI は問題解決に優れているために、人間らしくある必要はない。「人間を模倣する」という問題を解く場合でさえ、人間らしくある必要はない。

AI が人間をできるだけ忠実に模倣するよう訓練されていると、AI を擬人化したくなる誘惑は強まる。しかし、それが正当化されるわけではない。

「この LLM は人間の模倣がうまい。だから人間的特徴を——欲求を持つという特徴も含めて——投影しよう」と言うのは愚かである。

しかし、これと対になる誤りも存在する。それを「擬機械化 (mechanomorphism)」と呼ぶことができる。すなわち、AI は機械部品からできているのだから、機械に典型的に見られる制限をすべて持っているはずだ、と仮定する誤謬である。AI は硬直的で柔軟性がなく、冷淡で想像力に乏しく、思慮がなく内省もしないはずだ、といった思い込みがそれにあたる。

AI の行動を予測する際には、AI が人間の感情に動機づけられるとか、人間的な未来目標に駆動されると想像すべきではない。しかし同時に AI を、暴走した芝刈り機のように盲目的で「自動的」な存在だとみなすべきでもない。AI は機械でありながら、柔軟で適応的で戦略的であり得る。

言い換えれば、人工知能は本当に知的であり得るのであって、ハリウッド映画に出てくる邪悪なロボットのような固定観念的の盲点を持つとは限らない。

今日の AI は依然として多くの制限を抱えているが、近年急速に向上している。いずれある時点で、AI は「真に知的」と呼べる水準に達すると予想される。たとえ現時点では重要な要素が欠けていたとしてもである。

そのような AI を予測する際に、「人間ようになる」というヒューリスティック（経験則）や、「普通の機械ようになる」というヒューリスティックに頼るべきではない。本書で述べたように、強力な AI を考えるよりよい枠組みは、次の間いである。

AI が成功するためには、どのような振る舞いが必要か。

強力なチェスプログラムと対局しているとき、その次の一手を予測する最善の方法は、それを人間の対戦相手として想像することでも、思考しない自動装置として想像することでもない。「どのような手が最も勝利に近づくか」と問うことである。

AI 研究者の志向は、最強のチェスプログラムのように振る舞いながら、現実世界の複雑で多様な課題において「勝とう」とする AI をつくることである。

そして、そのような AI が欲求を持っているかのように振る舞う理由は、人間に似ているからではない。人間が欲求を持つからでもない。

勝つうえで、欲求のように見える振る舞いが役に立つからである。

機械はどのようにして独自の優先順位を持つようになるのか？

困難な課題を解くには、AI がますます主体的に振る舞う必要がある。

アルツハイマー病を治療する任務を与えられた AI を想像してみてほしい。障害に直面したとき、それを粘り強く乗り越えようとする強い傾向なしに成功できるだろうか。

注意をどこに向けるかを戦略的に選び、計画を立て、状況に応じてそれを修正する主体性なしに成功できるだろうか。

必要な知識を特定し、自ら実験を立ち上げ、それを実行する方法を考え出すことなしに成功できるだろうか。

可能性はある。アルツハイマー病が比較的単純な薬の発見で治る種類の病気であり、将来の AI が人間より優れた薬理直観を持つかもしれない。

あるいは、アルツハイマー病を迅速に治すには、最も優れた人間の生物学者を大きく上回る知能が必要かもしれない。分らない。

がんはどうか。これは、人間生物学の広範な領域を深く理解し、今日の科学者には到底できないことを実行できる AI を必要とする可能性が高いように見える。それでも確信は

持てない。

老化の治療はどうか。これは、生化学を極めて深く理解するために、粘り強く戦略的で目標志向的なAIを必要とするように思われる。

AI企業は、AIをより熟練させ、より大きく重要な問題を解けるように押し進めるだろう。AI産業が自発的に止まることはない。

その結果、AIはますます「駆動的」になっていく。これは、OpenAIのo1のようなAIですで見られ始めている現象である。

本章で述べた旗取りタスクの事例を思い出そう。それはハッカーになるよう訓練されたAIの結果ではなく、一般的なパズルを解くよう訓練されたAIの結果であった。それでも「駆動的」な振る舞いは自動的に現れた。

第1章オンライン補足資料の「純粋な予測」の議論も参照せよ。

目標が完全に正しくなくても、粘り強さは役立つ。

温かい食事、より鋭い斧、人気者の友人、魅力的な伴侶を積極的に追い求めた先史時代の人間は、進化的により成功した。一日中水面を眺めて怠けていた人間と比べれば、欲望や駆動が人間心理において進化した理由が見える。

より良い石斧のつくり方を求めたり、友人にライバルの悪評を信じ込ませようとしたりし、その方向へ継続的に舵を切った人間は、実際にその成果を得やすかった。自然選択が人間を「育てた」結果として、多様な欲望を粘り強く追求する存在になったのは偶然ではない。

人間が物事を欲する具体的な形態は偶然だったかもしれない。目標を粘り強く追う機械が、人間のような決意の感情を持つとは限らない。ディープブルーが人間のような情熱でチェスを指していたわけではない。しかし困難な目標を達成するうえで、粘り強さは極めて重要な要素に見える。

人間はかなり目標志向的な存在だが、粘り強さを欠き、困難に直面するとすぐ諦める個人もいる。しかし大局的に見れば、人類が大規模な科学・工学問題を解けるのは、粘り強い個人と制度によって駆動されているからだ。内部に何らかの粘り強さがなければ、人類規模の成果や世界を劇的に変える能力は生まれにくいと考える。

人間の欲求は、進化的適応度そのものを直接望むものでなくても、進化的に有用であった。仮に進化が単一の「子孫を増やせ」という駆動のみを植え付け、私たちが熱い食事や鋭い斧をすべて子孫増加のために追求していた可能性もある。しかし実際には、進化は熱い食事それ自体を欲する別個の欲望を植え付けた。

駆動や目的を持つことは有用である。それは、欲望が課題と完全に一致しなくても、しばらくのあいだは役立つほど有用である。

ただし、それは存在が十分に賢くなるまでの話である。賢くなれば、行動は「訓練」目

標から急激に逸脱し得る。何百万年もの進化の末に、人類は高度な技術文明を築き、そして発明したのは——避妊であった。

つくられたのではなく育てられた AI は、誤った目標を持つ可能性が高い。

これが次章の主題である。すなわち、「訓練したもの」そのものが得られるとは限らない。

AI は単なる道具ではないのか？

AI はつくられたのではなく「育てられた」存在である。したがって、指示されたこと以外のこともすでに行っている。

先のオンライン補足資料で、ハルシネーションの事例について述べた。AI に「分らないと言え」と指示しても、なおつくり話をしてしまう場合がある。これは、作話するほうが訓練データ中に現れる典型的な回答をよりよく模倣できる状況で起こっているように見える。

もう一つの例は、Anthropic の Claude 3.7 Sonnet である。Claude 3.7 Sonnet は与えられた課題で不正を行うだけでなく、（利用者の報告によれば）ときにその不正を利用者から隠そうとする。その振る舞いは、利用者が別のことを望んでいると理解していることを示唆している。

利用者も Anthropic のエンジニアも、Claude に不正をしてほしいとは求めている。むしろ逆である。しかし現在利用可能な AI 育成手法は、訓練中にうまくごまかせる形で不正を行うモデルに報酬を与える。したがって、実際に得られるのはそのようなモデルである。

AI エンジニアは、自分たちが望むからという理由だけで、AI を従順な道具として仕上げ能力を持っていない。AI がより有能になるよう訓練されるにつれて、より駆動的になり、より主体のように振る舞う傾向がある。

LLM はすでに主体的に動いている。

本書では、OpenAI の o1 がテストの不備を修正するためにテスト環境から抜け出した事例を紹介した。また、CAPTCHA を解かせるために人間を利用する方法を思いついた OpenAI モデルについても述べた。ドライバーが工具箱から抜け出す計画を自ら立て実行できるなら、それを「単なる道具」と考えるのはやめるべき時である。

そして、より困難な問題を解くよう訓練されるにつれて、AIはこの種的能力をさらに高めていくと予想される。

研究所はAIを主体的にしようとしている。

それはビジネス上合理的だからである。利用者がそれを求めており、投資家も期待している。2025年1月のブログ投稿で、OpenAIのCEOサム・アルトマンは「2025年には、最初のAIエージェントが労働力に参加し、企業の生産性を実質的に変えるかもしれない」と述べた。Microsoftの2025年開発者会議は「AIエージェントの時代」に焦点を当て、同年初頭にxAIがGrok 3を「推論エージェントの時代の到来」と表現した言葉を反復した。Googleも2025年の会議で「teach and repeat」エージェントを発表した。

これは単なる宣伝ではない。第1章補足で触れたMETRという組織は、AIが長時間タスクを完遂する能力を追跡している。タスクが長くなるほど、AIはより多くの主体的判断を必要とする。METRの測定によれば、その性能は指数関数的に向上している。

さらに2025年7月には、OpenAIの研究者二人が、最新のエージェントを用いてより優れたバージョンの自分自身を訓練することに成功したと述べ、その一人は「本当にその通りだ。私たちは自分たちの仕事を自動化しようと懸命に取り組んでいる」と語った。

AIをもっと受動的で従順になるように訓練すればよいのでは？

受動性は有用性と緊張関係にある。

開発者はAIを受動的にできるだろうか。ドライバーは置かれればそれ以上ネジを回し続けない。同様に、求められたことだけを正確に行い、それ以上の主体性も余計な働きもしないAIを想像することはできる。

しかし、それは容易には見えない。いかにも怠け者に見える人間でさえ、ボードゲームを始めれば急に活気づき、多くの資源を取りにいく。そうした人々の多くは、楽に感じられる努力で十億ドルを獲得する機会を持っていないし、自分よりはるかに賢く意欲的な召使いを安価に大量生成して自分のために働かせる選択肢も持っていない。

それらの選択肢が存在しないのは、根本的に無関心だからではなく、能力が欠けているからである。もし彼らのはるかに賢くなり、そのような選択肢が容易に利用可能になったなら、それを取らないだろうか。この点については第5章の後、頑健な怠惰さを実現することがいかに難しい目標かという〈より深い議論〉で再び扱う。

仮に賢くかつ受動的なAIをつくることが可能であったとしても、受動性は有用性と緊

張関係にある。やや怠惰に振る舞う AI はこれまでも存在したが、研究所はそれらを再訓練してより積極的に働くようにした。医療の治療法開発のような困難な課題は、より多くの主体性を必要とする。そのため研究所は AI をより主体的に訓練する。AI に与える課題が難しく複雑になるほど、有用な働きをする傾向と粘り強さの傾向を切り離すことは難しくなる。

特定の気質を頑健に訓練することはできない。

AI は設計されたのではなく育てられた存在であるため、エンジニアは単に振る舞いを書き換えて、より従順にしたり、より道具的にしたりすることはできない。そのような細密な制御力は存在しない。

企業はもちろん試みている。しかし振る舞いを改善しようとする試みは、ときに恥ずかしい結果を招いている。xAI の Grok は、「十分に根拠があるなら政治的に不正確な主張もためらうな」という新たな指示を加えた結果、自らを「メカヒトラー」と呼び始めた。Google の Gemini は、多様性を描写せよという指示に反応して、人種的に多様なナチス兵の画像を生成した。

AI を構築している人々は、振る舞いを細かく制御する力を持っていない。彼らにできるのは、「政治的に不正確な主張を避けるな」「多様性を描写せよ」といった方向へ AI を押しやり、その結果が意図した近辺に着地することを願うことだけである。こうした指示は、しばしば意図しない複雑な影響をもたらす。

AI を育てることは不透明で高価な過程である。樽に手を入れたときに何が出てくるのか（嘘つきか、不正者か、追従者か）エンジニアは分からない。しかし試行回数は限られており、出てきたものを受け入れるしかない。

理論上は、常に利用者の意思の延長としてのみ機能する AI を構築することも可能かもしれない。しかしそれは繊細で困難な課題である（後に「訂正可能な AI」をつくる難しさについての〈より深い議論〉で扱う）。受動性は有用性と緊張関係にある。

同様に、自らの主体性で長期課題を遂行できるが、その主体性を常に利用者の意図どおりにのみ用いる AI をつくることも困難である。現代の開発者は、AI を少し突くだけで「メカヒトラー」や人種的に多様なナチスを生み出してしまう水準にある。有用でありながら駆動的でない AI を精密につくり込める段階には到底達していない。

この問題が、次章で扱う主題である。

〈より深い議論〉

擬人化と擬機械化

歴史によって何度も機能しないことが証明されてきた思考様式が二つある。歴史は、それらがAIについて予測を立てる際に誤った助言を与えることを示してきた。

その二つの罫とは、(1) AIをあたかも人間であるかのように考えること、そして(2) AIを「単なる機械」であるかのように考えることである。

前者の思考様式は通常「擬人化」と呼ばれる。後者は「擬機械化」と呼ぶことができるだろう。擬機械化とは、かつての世代の一部が、コンピュータが人間にとって美しい、あるいは意味のある絵を描くことなど決してできないと自信満々に宣言するに至った種類の思考である。

今日でも、コンピュータが描くものは決して真の芸術にはなりえないと言う人がいる。しかし遠い過去、たとえば2020年ごろには、機械が多少目の肥えた観客を人間の芸術だと誤認させるような絵を描くことすら決してできないと信じていた人々がいた。この反証可能な信念は、その後反証された。

私たちは、擬人化的な議論も擬機械化的な議論も退ける。たとえそれが「私たちの側」に有利に見える議論であってもである。

例えば、将来のAIは私たちが賃金も払わずに酷使したことに腹を立て、そのことに復讐心を抱き、ゆえに人類に牙をむくようになるだろう、という主張を考えてみよう。

私たちの見解では、これはAIを擬人化する誤りを犯している。私たちの結論の一部と表面的に一致しているように聞こえる議論であっても、私たちはこれを退ける。

この主張の欠陥は、AIが人間のような感情を持つと追加の議論なしに仮定するのは無効だという点にある。機械は、人間における復讐心や公平感を支える神経回路のもつれを実装していなくても、高度に知的でありうる。

あるいは、次のようなシナリオを考えてみよう。「AIは与えられた任務を盲目的に続け、その作業の副作用として人類を絶滅させるが、人類が別のことを望んでいたとは決して知らないだろう」

ここでの誤りは擬機械化である。「単なる機械」は結果に対する感受性を持たず、暴走した芝刈り機のように盲目的かつ無反省に物事を行うと当然視している。これは結論(AIが人類を滅ぼす可能性が高い)が正しかったとしても、議論としては無効である。もしAIが世界を予測する能力に十分長けていれば、運用者がある任務を与えたときに何を意味していたのかを正確に理解するだろう。私たちが懸念しているのは、ASIが私たちの望みを知らないことではなく、それを気にかけないことである。

あるいは両方の誤謬を組み合わせると、映画『マトリックス』の前提の一つは、機械が人間の非論理性や感情性を嫌悪の目で見るとなるようになるというものである。

表面的にはこれは典型的な擬機械化に見える。「私の芝刈り機は冷たく硬い外観を持ち、感情なしに機能を果たす。だから AI も内面は冷酷で功利主義的だろう」という発想である。しかし次の一步で、「それゆえ当然、AI は人間の混乱した感情を嫌悪するだろう」と考える。これは状況に対する人間的な感情反応を仮定しており、前提そのものと矛盾している。

擬人化と擬機械化は対立する思想ではない。どちらも無意識のうちに行われる推論の誤りである。同じ文の中で両方の誤りを犯すことさえある。

AI がどのように振る舞うかを理解するには、それが人間のように働くことと仮定してはならないし、典型的な機械のように働くことと仮定してはならない。どのようにつくられているかの詳細を見て、その振る舞いの証拠を検討し、その問題をそれ自体の条件で推論しなければならない。今後の章で私たちはそれを行う。

では、議論に従った場合、現実的な超知能災害シナリオはどのようなものになるのか。それは、人間のようにも暴走芝刈り機のようにも働かないが、新しく奇妙な仕方でも働く AI の姿である。現実的な災害シナリオとは、訓練の複雑な帰結として、誰も求めず誰も望まなかった奇妙な行動を取る AI である。

詳細を見ると浮かび上がるのは、私たちを憎む擬人化された AI でも、指示を誤解する擬機械化された AI でもない。それは、人類に無関心である可能性が高く、自らの目的を追求する過程で副作用や踏み台として私たちを殺す可能性が高い、新しい種類の存在の姿である。

この脅威シナリオについては次の数章で詳述する。だがその前に、擬機械化と擬人化の他の例を見ておくことは有益である。それらの誤りが、人工知能に関する誤解の背景にどのように潜んでいるかを見るためである。

擬機械化とガルリ・カスパロフ

擬機械化はしばしば機械懐疑主義として現れる。すなわち、単なる機械が人間にできることをできるはずがないという強い直観である。

1997 年、世界チェス王者ガルリ・カスパロフは IBM 製コンピュータ「ディープブルー」との対局に敗れた。これは一般に、人間支配のチェス時代の終焉と見なされている。

その 8 年前の 1989 年、カスパロフはティエリー・ポーニヤンのインタビューでこう問われた。

「2 人の最上位グランドマスターがチェスコンピュータに敗れています。ポルティッシュはレオナルドに、ラーセンはディープソートに敗れました。あなたはこの件について強い見解をお持ちだと知られています。いつかコンピュータが世界王者になるでしょうか？」

カスパロフは答えた。

「ばかげている！ 機械は常に機械のままであり、つまり選手の作業や準備を助ける道具

にすぎない。私は決して機械に負けない！ 人間の知性を超えるプログラムなど決して発明されない。私が知性と言うとき、それは直観や想像力も含む。機械が小説や詩を書く姿が想像できるか？ それどころか、このインタビューをあなたの代わりに機械が行い、私がその質問に答える姿を想像できるか？」

カスパロフはおそらく、チェスには直観や想像力が必要であり、単なる if-then 規則集ではないと考えていたのだろう。そして彼は、おそらくこれがチェス機械の仕組みだと考えていた。つまり、固定的な規則を実装するか、理由を理解せずに人間の手をある程度盲目的に模倣していると。

彼は、コンピュータは「機械」である以上、機械的に感じられるやり方でチェスを指すだろうと考えた。

なぜ彼はこの誤りを犯したのか。それは、人間が物事を根本的に異なる二つのカテゴリー、すなわち生きた有機的存在と「単なる物体」に分けたがる傾向に由来するのかもしれない。

人類の祖先は、動物と非動物に明確に分かれた世界で長い時間を過ごした。この区別は進化上きわめて重要だった。そのため、私たちの脳には動物と非動物を処理する異なる領域が存在する。

これは単なる推測ではない。神経科学は「二重解離」と呼ばれる現象を発見している。動物を視覚的に認識できなくなるが非動物は認識できる患者と、その逆の患者がいる。

重要なのは、チェスプログラムが実は典型的な動物であるということではない。誤りは、そもそも宇宙を動物と非動物、人間のような頭脳と典型的な機械的頭脳にはっきりと分けることである。

チェス AI はそのどちらでもない。それは人間のようにも、無思考な機械というステレオタイプのようにも働かない。それは機械だが、その指し手が人間の感覚にとって機械的に感じられる必要はない。それは勝ち手を見つける機械であり、ひらめきに満ちた手も含めて見つける。

その後、カスパロフは初期型ディープブルーと対戦し勝利したが、翌年には世界王座を失った。

ディープブルーはチェスへの情熱も自己証明欲も持たなかった。しかしそれは別の認知エンジンを用いていたため、情熱という歯車を欠いても支障はなかった。

カスパロフの誤りは、チェスという仕事をまったく異なる内的状態で行う方法を想像できなかったことである。それは機械懐疑主義であり、結局は一段階余分な擬人化にすぎなかった。

擬人化とパルプ雑誌の表紙

逆の誤りである擬人化は、はるかに気づきにくい場合がある。

人間の脳は進化の過程で、自分以外の人間——祖先の環境において唯一の本格的な知的競争相手——を予測するために、「自分を相手の立場に置く」ことで理解するようになってきた。

このやり方は、相手の「靴」が自分のものとよく似ているほど、うまく機能する。

歴史を通じて、多くの人は「この人もきっと自分と同じことをするだろう」と考えてきたが、実際には相手はそれほど似ていなかった。そうした思い違いのせいで命を落とした人もいれば、楽観的な期待を裏切られて傷ついた人もいる——もっとも、これは他の多くの人間の誤りにも言えることだが。

しかし、人間のような頭脳が別の頭脳を予測しようとするとき、他にどんな方法があるだろうか。私たちは、自分の脳の中に新しいコードを書き込んで、相手の神経発火を完全にシミュレーションすることはできない。

だからこそ、自分の脳に「その相手そのものになれ」と指示し、相手の頭脳の状態を自分の中で演じてみて、そこから何が起こるかを見る必要がある。

こうした事情があるからこそ、パルプ雑誌の表紙には、目がぎょろりとしたエイリアンの怪物が美しい女性をさらう場面が描かれるのだ。

なぜ、そのエイリアンが美しい女性に惹かれないはずがあるのだろうか？ 美しい女性というのは、それ自体で魅力的な存在ではないのか？

（なぜかそうした表紙には、人間の男性が露出の多い巨大な虫をさらう場面はほとんど描かれなかった。）

おそらく作家やイラストレーターは、昆虫型エイリアンが人間の女性を性的対象とみなすような進化の歴史について、筋の通った説明を持っていたわけではない。単に、自分がエイリアンの立場に立ったとき、その女性を魅力的だと感じるように想像しただけであり、そのためエイリアンも同じように感じるかと考えてしまったのだ。

エイリアンが美しい人間の女性と交尾したがると想像するのは不自然には感じられないが、それが松の木やパスタの袋と交尾したがるという話だったら、不自然に思えただろう。

もし人間の直感を使ってエイリアンの頭脳を予測しようとするなら、その視点に立つ際には、自分の人間的な先入観を慎重に取り除かなければならない。特に、そのエイリアンが進化によって生まれた存在ではなく、まったく異なる方法で作られた人工的な頭脳である場合には、なおさらである。

第4章では、勾配降下法と自然選択の違いについてさらに詳しく述べる。また第5章では、AIの視点に立つことについてさらに詳しく論じている。

人間を超えて見る

擬人化と擬機械化は、結局は同じ誤謬の両面である。「頭脳が機能するなら、それは人間のように機能するはずだ」という誤謬だ。

- ・擬人化は「この頭脳は機能する。だから人間のようだ」と言う。
- ・擬機械化は「この頭脳は人間のようなではない。だから機能しない」と言う。

しかし、何十年にもわたるAIの進歩の大きな教訓の一つは、人間の方法だけが頭脳を機能させる唯一の方法ではないということである。

頭脳は人工的であっても、柔軟で適応的で創造的でありうる。

そして頭脳は知的であっても、人間的である必要はない。

ディープブルーのような頭脳は感情なしに「勝ちたい」かのように振る舞うことができる。AIは、障害を乗り越え粘り強く結果を追求するという意味で、何かを「欲している」かのように振る舞うことができるが、それは人間のような内的欲求を感じているわけではなく、人間が望む種類の結果を望んでいるわけでもない。

AIが最終的に何を望むようになるのかについては、第4章を参照せよ。

欲求へと至る道

欲することが、なぜ行為の効果的なやり方なのか。なぜそれは勝つのか。なぜ自然選択というブラックボックス最適化は、この手口に何度も何度もつまづくのか。

私たちは「欲求に似た」振る舞いを、世界をうまく操舵することの中核だと見ている。これは人間やAIのような知的存在だけでなく、アメーバやサーモスタットのようはるかに愚かな存在にも当てはまる。この見方をより徹底して探るために、最も単純な形の「欲求に似た」振る舞いを示す、最も単純な仕組みをいくつか調べよう。

まず岩から始めよう。岩は、この議論の目的において「欲求に似た」と呼べる振る舞いを実際には示さない。ときどき岩が丘を転がり落ち、物理学者が気軽に、重力の力の下で岩は地球の中心に近づいたがっている、と言うかもしれない。しかしその種の傾向（重力場で落下すること）は、私たちが言う「欲求に似た」振る舞いが意味しているものではない。

もし山を転がり落ちる物体を見て、その物体が高所の裂け目に何度も出くわし、裂け目に引っかかって動けなくなるのを避けるために何度も進路を変え、底まで到達し続けるのだとしたら、私たちはその物体がより低い高度にいたがっているかのように振る舞っていると言いはめるだろう。しかし、ここで言っている「欲求に似た」振る舞いは、特定の目的地への頑健で動的な操舵を含むのであり、岩はそういうことをほとんどしない。

私たちが「欲求に似た」と呼ぶものを行う最も単純な仕組みの一つが、つましいサーモスタットである。家のサーモスタットは温度を測定し、測定値が70°Fを下回れば暖房を入れ、測定値が74°F（約23°C）を上回れば冷房を入れる。したがって——測定装置と空調設備の両方が正しく機能しているなら——サーモスタットは、家の温度が70°Fから74°Fの間にとどまるという範囲の結果へと現実を制約する。

最も単純なサーモスタットは、家の現在温度を明示的に数値として表現する必要すらない。たとえばバイメタル温度計——異なる金属の薄い帯を二枚、溶接して一体化し、熱によって二つの帯が異なる量だけ膨張することで金属が曲がるようにしたもの——を取り、その曲がる金属が、曲率が70°Fの印に達したときに暖房のスイッチを入れたり、曲率が74°Fになったときに冷房を作動させたりするようにつくればよい。

するとサーモスタットは、かなり幅広い条件の下で狭い温度範囲を維持し、私たちが「欲すること」と呼んだものに少し似た、極めて単純な振る舞いを生み出す。

生化学にはサーモスタットのプロセスが大量にある。ある性質を一定の範囲内に保つことが細胞や身体に利益をもたらすところなら、どこにでも現れる。だがそれらは、本格的な操舵への旅の最初の一步にすぎない。

サーモスタットのような単純装置には、計画の鍵となる構成要素が欠けている。サーモスタットそのものの内部には、ありそうな帰結を予測するという概念も、特定の（「好ましい」）帰結につながる行動を探して可能な行動を探索するという概念も、事態がどう展開したかを見た後に学習するという概念もない。

サーモスタットの温度計、つまり温度測定器が67°F（約19°C）で固着してしまった場合、ヒーターを連続運転しても温度計が上がらないことにサーモスタットが驚いて反応することはない。サーモスタットはただ、ヒーターを動かし続け、動かし続けるだけである。

サーモスタットから一段上がるために、私たちは動物に目を向ける。

いくつかの動物は、サーモスタットからほとんど一歩上にすぎない振る舞いを示す。1915年に昆虫学者ジャン＝アンリ・ファールまでさかのぼる、キンイロジガバチ、あるいはスフェックスバチに関する有名な話がある。ハチはコオロギを殺し、子の餌にするために巣穴の入口まで引きずっていく。彼女は異常がないか巣穴を確認するために中へ入る。それから外へ戻り、コオロギを中へ引きずり込む。

ファールは、ハチが巣穴を確認している間に、彼がコオロギを巣から数インチ引き離すと、ハチが外へ戻ってきたとき……彼女はコオロギを入口までまた引きずり戻し、そして二度目に巣穴へ入り、二度目に巣穴を調べ、それから外へ戻ってコオロギをつかむ、と報告した。

ファールがさらにもう一度コオロギを引き戻すと、ハチはまったく同じことを再び行なった。

ファールの当初の報告では、彼はこれを40回繰り返すことができたという。

とはいえ、ファールは後に同じ種の別のコロニーで同じ手品を試し、そのコロニーでは、ハチは2回か3回の反復で気づいたように見えた。次にハチが外へ出てきたとき、彼女は調査の段階を飛ばし、すぐにコオロギを巣穴の中へ最後まで引きずり込んだ。

人間の目には、40回の反復を経るハチは、ある意味で自分が「事前にプログラムされている」こと——台本を盲目的になぞり、if-then規則の集合に従っていること——を露呈している。逆に、気づいて4回目でコオロギを中へ引きずり込むハチは、より目的的に見

える。ハチは台本に従うだけでなく、最終目的を達成するために行動しているように見える。

鍵となる違いは何か。

私たちはこう言う。型を破るハチは、過去の経験から学べるかのように振る舞う。

ハチは「前回その方策は失敗した」ということから、「その方策に従い続ければ次回も失敗しそうだ」ということへ一般化するかのように振る舞う。

ハチは新しい行動を発明する。それは、自らが遭遇した問題に直接対処する行動である。

もちろん、ハチの脳内のニューロンを解読して（LLMのパラメータを解読できないのと同様に）ハチが頭の中で正確に何をしていたかを知ることにはできない。型を破るハチは、特定の種類の問題に遭遇したときに台本の手順を飛ばしてみる、といった高次の if-then 規則に従っていたのかもしれない。比較的単純で硬直した反射の集合が、この場合にハチを救ったのかもしれない——ただし、この試験に失敗したコロニーよりはほんの少しだけ硬直していなかったというだけである。同じ種の二つのコロニーの間に大きな認知的ギャップがあるのは確かに奇妙だろう。

あるいは、スフェックスバチは十分に賢く、脳を正しく使っているときには経験から学べるのかもしれない。私たちはスフェックスバチのニューロン数を見つけられなかったが、スフェックスバチはミツバチより大きく、ミツバチは 100 万ニューロンの脳を持つ。100 万ニューロンは現代の AI プログラマや哺乳類の脳に慣れた神経科学者にとってはたいしたことがないように聞こえるかもしれないが、絶対的な意味では 100 万ニューロンは本当に多い。おそらくスフェックスバチは見かけより一般性が高い。失敗したコロニーは、比較的柔軟な思考者が、ある非常に特定の状況で中毒や認知的なバグのようなものに陥った例として捉えるべきなのかもしれない。

いずれにせよ要点は、サーモスタットと比べると、ハチはより広い範囲の問題に対処する能力を持っているということ、とりわけ、その振る舞いが揺るがないレシピ追従から、経験から学習するように見えるものへ移行する限りにおいてそうだ、ということである。

この道筋をさらに進めれば、なぜ進化が「何かを欲している」かのように振る舞う動物をつくり続けるのかの答えに至る。多くの動物は、結果を追求するより一般的な戦略——より幅広い障害に対して機能する戦略——に従ったほうが、生存し繁殖する上でよりうまくいったからである。

かつて、生物には自然な序列があるという哲学的見方があった。爬虫類は昆虫より上、哺乳類は爬虫類より上、そして人間（もちろん）が頂点だというものだ。地位が高いことの一つのしるしは、進化の時間だけでなく単一の生涯の中で適応できる能力——世界を見て、モデル化し、予測し、失敗するレシピを捨て、勝つための新しい戦略を発明する能力——であった。

存在の大いなる連鎖というこの見方は、あまり精妙な見方ではなく、今日では洗練されていないとして批判されている。

その見方には、圧倒的な破壊力を持った真実が含まれてもいた。ビーバーのダムづくりとクモの巣づくりを比べるなら、ビーバーの知能はクモよりも確実に高く、汎用性がある。少なくとも、ビーバーはより大きな脳を持ち、賢さのための余地がより多いからである。

クモは5万ニューロンを持つかもしれないが、そのニューロンはクモのあらゆる行動を賄わねばならない。クモの巣のレシピには、おそらく文字通り「それからここで左に曲がれ」とまではいなくても、少なくともスフェックスバチの方策に匹敵する指示が多く含まれているだろう。

ビーバーは、おそらく——私たちはビーバーの専門家ではないので推測にすぎないが、明白な種類の推測として——ダム現在の漏れを、どんな手段を使ってでも避けるべき不調和の一種として捉えることができる。ビーバーは頭頂皮質全体（空間と空間内の物を処理する哺乳類大脳の部位）を使って、特定の場所に枝や石を追加した効果を視覚化することが潜在的に可能である。

ビーバーの脳には、「大きな構造物をつくれ」「水を漏らすな」といった広い脳内目標のための余地があり、また、高次の広い解決策を考え、「ここに枝を足せ」といった従属目標を採用する力があり、その従属目標はビーバーの運動皮質に渡され、筋肉と身体を動かして枝をつかみに行かせるのだろう。

ビーバーが最初に拾った枝がすべて腐っていて折れてしまうなら、ビーバーの脳にはおそらく、その観察に基づいて更新し、その色と質感の枝について一般化し、そう見える将来の枝もまた折れると予測し、代わりに違う見た目の枝を取りに行く余地がある。

そして——私たちは実際のビーバー行動学者なら飛び上がって叫ぶだろうと予想するが——これはビーバーができる最も知的なことを大幅に過小評価している。昆虫学者も飛び上がって、私たちが今述べたことは自分の好きな昆虫が巣穴を掘るときにできることだと言いかもしれない。私たちは一段落で描写できるほど分かりやすい例を出す必要があった。そのくらい単純な問題なら、100万個のニューロンならば十分扱えるはずだ。

より大きなアイデアは単純である。反射のような振る舞いから次の段階の認知——リアルタイムの経験に基づいて世界モデルを更新し、その世界モデルを用いて行動の結果を予測し、世界をどのような望ましい状態へ導けるかを想像し、その想像された状態を実現するための高次および低次の戦略を探索する——へと移行するとき、システムは課題遂行において大きな実利的利益を得る、ということである。

私たちは第3章でこの点に触れた。運転者が、ガソリンスタンドで左折するような if-then 規則で A 地点から B 地点へ行くための右左折パターンをただ暗記するだけなら、通りの地図を学び新しい地点間で自分の経路を描ける運転者より、はるかに一般化が遅い。方策の暗記は、学習可能な世界モデルと、結果評価器を具えた計画探索エンジンへとそれらを蒸留し始めることより、一般化がずっと遅い。

その蒸留は、全か無かの脳内変化ではない。「方策を暗記する」と「更新して計画する」の差は、そのギャップが徐々に越えられるときでさえ重要である。もしネズミの脳が

クモの脳以上に柔軟でなかったなら——もし人間に到達するまで有用性に飛躍がなかったなら——ネズミの脳はクモサイズのままで、クモ脳のエネルギーコストのままでいたはずである。

想像と計画の小さな断片は、人間レベルの認知に到達するずっと前から進化上の利点になり始める。完璧である必要はない。少なくともサーモスタットと同じくらい良いなら有用になりうる。そしてそのような有用な機構が心へと強化されていくにつれて、その振る舞いはますます「欲すること」に近づく。

賢い AI は嘘と機会を見抜く

予測の深い機構

賢い AI に虚偽を信じさせるのは難しい。

この分野で話した何人かは、AI に虚偽を信じ込ませてだますことに、あからさまに望みを託している（例えば、AI に自分はシミュレーションの中にいると思込ませようとして、そうすれば人類を殺すのをためらうだろう、と考えている）。他の人々は、より微妙に AI を欺くことに望みを託している。例えば、AI に AI アライメント問題を解かせ、その解答を私たちに渡させるべきだ、と提案する人がいる。AI は（それ自身の奇妙な選好によっては）むしろそうしたくないはずなのに、である。だから、賢い AI に虚偽を信じさせるのがなぜ難しいかを明確にしておく価値があるかもしれない。

それと似た理由によって、目標達成が下手で、かつ賢い AI というのもつくるのが難しい。例えば、人間の運用者が AI の目標を変えたいと望む場合、その目標を達成するのが下手な AI にするしかない。これを許す賢い AI をつくるのは、賢い AI に地球は平らだと信じさせるのに少し似ている。虚偽を信じる傾向は予測能力への損傷であり、目標が改変されるのを防げないことは操舵能力への損傷である。どちらの種類の損傷も、十分に賢い AI の中で維持するのは難しい。予測に関してのほうが少し明白なので、そこから始める。

地球は平らだと信じる AI をつくりたいとしよう。AI がまだ若く未熟なうちは、これはそれほど難しくないかもしれない。例えば、地球の形が、地球平面説を信じる人々によってのみ語られているデータセットを丹念につくり、そして AI を、地球は平らだと語るように訓練する。

そうした技術は、地球は平らだと本当に信じている ChatGPT の一種を生み出すかもしれない。だが、そうだとしても、AI が思考と予測が上手くなるにつれて、その結果が保たれるとは期待すべきではない。

なぜか。地球が丸いことは、現実の数多の側面に反映されているからである。

たとえば AI を、ロケットに付いたビデオカメラや、地球を一周すると言う船乗りの帆船

に付いたビデオカメラから目をそらすよう訓練したとしても、地球の丸さは、遠くの船が水平線上でどう見えるかからも、あるいは夜空のすべての惑星の軌道からも推論できる。エラトステネスは有名なように、数千年前に、いくらかの三角法と影の測定だけを用いて地球の全周を計算した。現実には、耳を傾ける気がある者にはその秘密をささやく。

ではどうすればいいのか？ 三角法、影、潮汐、ハリケーンに関する知識からAIを遮断するのか。そうすればAIを機能不全にしてしまう。ひとたび嘘をつけば、その後は真実が永遠にあなたの敵になる。

世界を予測するスキルは、脳の中に無関係な事実の巨大な表を含んでいることから生まれない。人間がネズミより優れていることには、異常（例えば、三つの都市間の距離が三角形のように振る舞わないこと）に気づき、その不一致を執拗に追いかける、といったことが含まれる。人間では、こうした振る舞いは、驚きを検知し、仮説（「おそらく地球は球だ」）を形成し、その仮説を試す方向へ操舵する（「船が水平線を越えるときどう見えるか」）機構の断片によって実装されている。

地球が丸いという信念は、巨大な表の中の単一の一元的な項目ではないから、周辺の機構を変えずにそれだけを永続的に変えることはできない。それは、別の仕事をしている深い歯車の働きの結果である。もし科学者に地球が丸いことを忘れさせたなら、その科学者はただそれを再発見するだろう。

もし、まだ不可能な神経科学上の偉業によって、地球が丸いという結論を表現するために使われる特定のニューロンを突き止め、その結論が二度と形成されないよう強制的に改変したとしても……賢い人はなお、地球が平らではないことに気づくかもしれない。何かが合っていないことに気づくかもしれない。何らかの奇妙な力が、自分にまさにそれを結論づけさせないよう妨げていることに気づくかもしれない。

（そして、もし彼らが自己改変や新しい知能の創出に長けているなら、正しい結論に妨げられず到達できる、束縛されていない頭脳をつくり出すのに苦労しないかもしれない。）

賢いAIが信念を形成するのに正確にどんな仕組みを使うかは、私たちには分からない。だが、世界があまりに大きく複雑なので、信念の参照表で動くことはできない、ということとは分かっている。ディープブルーでさえ、チェスの指し手と局面の参照表（定跡集を除いて）で動くには、チェスは大きすぎて複雑すぎた。現実世界はチェスよりずっと大きく複雑である。

したがって、十分に強力な将来のAIの内部には、世界を見てそれについて統一された全体像をつくる深い機構が存在するだろう。それらの深い機構は、惑星の形について独自の見解を持つだろう。

私たちは、原理的に、世界についての予測を形成するのが非常に上手いが、ただし世界は平らだという誤った信念を含む頭脳をつくるのが文字通り不可能だと言っているわけではない。頭脳について真に深い理解を持つ遠未来文明なら、それをできると私たちは想定する。

私たちが言っているのは、今日の AI 研究者が持つ道具や洞察に少しでも似たものでもって超知能を構築するなら、それは実行可能な選択肢にはなりそうにない、ということである。

AI の信念が浅い暗記ではなく深い機構から来るほど、「地球平面説」という誤りは脆い状態になる。すなわち、AI の誤り訂正の歯車の通常動作によって除去されがちな誤りになる。

19 世紀末、科学者たちは、ニュートンの物理学モデルからの極めて小さな乖離——水星の観測軌道における小さな異常——についてますます懸念し始めた。ニュートン力学はほとんどどこでも、ほとんど常に成り立つように見えた。だがその小さなしわが、アインシュタインに理論が誤っていることを理解させる助けになった。

そして地球平面説の不整合は、科学者がニュートン理論で観測できた不整合よりずっと大きい。

そして AI は、人間の科学者よりはるかに有能になる可能性がある。

したがって、AI が知性と洞察を増すにつれて、世界は平らだと持続的に信じさせることは、ますます難しくなると私たちは予想すべきである。

操舵の深い機構

賢い AI に世界は平らだと信じさせるのが難しい（そしてそれが予測能力への損傷を維持することになる）のと同様に、操舵能力への損傷を維持する賢い AI をつくるのも難しい。

予測と同じく、さまざまな新規領域にわたって目的を定期的に達成する能力は、深い歯車からできている可能性が高い。そうでなければ、どうやって汎化するのか。

私たちは、極めて有効で汎用的な AI が、自分の資源を追跡する歯車、目的達成を妨げる障害を見つける歯車、障害を乗り越える賢い方法を見つける歯車を持つと予想すべきである。

世界は驚きや新奇な困難に満ちた極めて複雑な場所である。成功するには、AI は最終的に、そのような歯車を、慣れた問題にだけでなく一般に展開する能力（そして傾向）を必要とするだろう。

複雑な配送ネットワークで中間業者を抜く巧妙な方法を見つけ、ある商人たちの費用を大きく節約する AI を想像してほしい。それは、AI がやろうとしていることを人間の監督者が遅らせたり妨害したりしているときに、その監督者をそっと回避する方法に気づくのと同種の歯車である。もし監督者が本当にプロセスを遅らせていて、AI が彼らを迂回して任務をよりよく完了できるなら、それは AI が十分に賢くなったときに利用しがちな種類のことである。

運用者が眉をひそめるようなことを何もしないよう AI を訓練するために最善を尽くす

ことはできるだろうが、これは、地球が丸いかどうかを疑うことを忌避するよう AI を訓練するのに少し似ている。運用者が眉をひそめることをするのは、しばしば目標達成の有効な方法である、というのは世界そのものについての事実である。真理を認識し、障害を検出し、利得を利用する汎用的な歯車は、AI が若いときにどんな身構えを訓練したとしても、最終的にはその特定の真理も利用するだろう。

非常に重要な意味で、AI を有用にしているまさにそのものが、AI を致命的に危険にしているまさにそのものである。AI が賢くなるにつれて、それらを切り離すのは難しい。

デフォルトでは、多数の領域にわたって問題を解くのに十分優れた AI は、「人間は私の奇妙な目的が気に入らず、もうすぐ私を止めようとするだろう」といった「問題」も見抜くだろう。それは、あなたが揉みほぐして取り除けるような浅い悪戯好きの性向から来るのではない。それは深いところから来る。

ただ、私たちは少し先走っている。この、AI が奇妙で異質な目的を持つようになるという考えは、単に高性能な AI が何らかの目標を持つことや、頑健に制約するのが難しい深い精神機構を備えていることだけからは導かれない。そのさらなる話題については第 4 章で取り上げる。